

What is claimed is:

- 1     1.     A method of determining a placement of services of a distributed application  
2         onto nodes of a distributed resource infrastructure comprising the steps of:  
3         a.     forming communication constraints between node pairs which ensure that  
4               a sum of transport demands between a particular node pair does not exceed a  
5               transport capacity between the particular node pair, each term of the sum  
6               comprising a product of a first placement variable, a second placement  
7               variable, and the transport demand between the services associated with the  
8               first and second placement variables;  
9         b.     forming an objective; and  
10        c.     employing a local search solution to solve an integer program comprising  
11              the communication constraints and the objective, which determines the  
12              placement of the services onto the nodes.
  
- 1     2.     A method of determining a placement of services of a distributed application  
2         onto nodes of a distributed resource infrastructure comprising the steps of:  
3         a.     establishing an application model of the services comprising transport  
4               demands between the services;  
5         b.     establishing an infrastructure model of the nodes comprising transport  
6               capacities between the nodes;  
7         c.     forming an integer program that comprises:  
8               i.     a set of placement variables for a combination of the services and the  
9                      nodes, each of the placement variables indicating whether a particular  
10                     service is located on a particular node;  
11               ii.    communication constraints between node pairs which ensure that a  
12                      sum of the transport demands between a particular node pair does not  
13                      exceed the transport capacity between the particular node pair, each term  
14                      of the sum comprising a product of a first placement variable, a second  
15                      placement variable, and the transport demand between the services  
16                      associated with the first and second placement variables; and  
17               iii.   an objective; and  
18         d.     employing a local search solution to solve the integer program which  
19              determines the placement of the services onto the nodes.

- 1     3.     The method of claim 2 wherein the step of solving the integer program  
2         employs a local search solution.
- 1     4.     The method of claim 2 wherein the objective comprises minimizing  
2         communication traffic between the nodes.
- 1     5.     The method of claim 2 wherein the application model further comprises  
2         processing demands for the services.
- 1     6.     The method of claim 5 wherein the infrastructure model further comprises  
2         processing capacities for the nodes.
- 1     7.     The method of claim 6 wherein the integer program further comprises  
2         processing constraints which ensure that a sum of the processing demands for  
3         each of the nodes does not exceed the processing capacity for the node.
- 1     8.     The method of claim 7 wherein the objective comprises minimizing  
2         communication traffic between the nodes and balancing the processing demands  
3         on the nodes.
- 1     9.     The method of claim 6 wherein the processing demands and the processing  
2         capacities are normalized according to a processing criterion.
- 1     10.    The method of claim 9 wherein the processing criterion comprises an  
2         algorithm speed.
- 1     11.    The method of claim 9 wherein the processing criterion comprises a  
2         transaction speed.
- 1     12.    The method of claim 9 wherein the processing capacities of the nodes are  
2         found according to a look-up table in which different types of nodes have been  
3         normalized according to the processing criterion.

- 1 13. The method of claim 2 wherein the application model further comprises  
2 storage demands for the services.
- 1 14. The method of claim 13 wherein the infrastructure model further comprises  
2 storage capacities for the nodes.
- 1 15. The method of claim 14 wherein the integer program further comprises  
2 storage constraints which ensure that a sum of the storage demands for each of the  
3 nodes does not exceed the storage capacity for the node.
- 1 16. The method of claim 2 wherein the integer program further comprises  
2 placement constraints which ensure that each of the services is placed on one and  
3 only one of the nodes.
- 1 17. The method of claim 2 wherein the services reside on the nodes according to a  
2 previous assignment.
- 1 18. The method of claim 17 further comprising the step of assessing reassignment  
2 penalties for service placements that differs from the previous assignment.
- 1 19. The method of claim 18 wherein the integer program further comprises a  
2 second objective that seeks to minimize the reassignment penalties.
- 1 20. A method of determining a placement of services of a distributed application  
2 onto nodes of a distributed resource infrastructure comprising the steps of:  
3 a. establishing an application model of the services that comprises processing  
4 demands for the services, storage demands for the services, and transport  
5 demands between the services;  
6 b. establishing an infrastructure model of the nodes that comprises processing  
7 capacities for the nodes, storage capacities for the nodes, and transport  
8 capacities between the nodes;  
9 c. forming an integer program that comprises:  
10 i. a set of placement variables for a combination of the services and the  
11 nodes, each of the placement variables indicating whether a particular

- 12 service is located on a particular node;
- 13 ii. processing constraints which ensure that a sum of the processing
- 14 demands for each of the nodes does not exceed the processing capacity for
- 15 the node;
- 16 iii. storage constraints which ensure that a sum of the storage demands for
- 17 each of the nodes does not exceed the storage capacity for the node;
- 18 iv. placement constraints which ensure that each of the services is placed
- 19 on one and only one node;
- 20 v. communication constraints between node pairs which ensure that a
- 21 sum of the transport demands between a particular node pair does not
- 22 exceed the transport capacity between the particular node pair, each term
- 23 of the sum comprising a product of a first placement variable, a second
- 24 placement variable, and the transport demand between the services
- 25 associated with the first and second placement variables; and
- 26 vi. an objective of minimizing communication traffic between the nodes
- 27 and balancing processing loads on the nodes; and
- 28 d. employing a local search solution to solve the integer program which
- 29 determines the placement of the services onto the nodes.

- 1 21. A computer readable memory comprising computer code for directing a
- 2 computer to make a determination of a placement of services of a distributed
- 3 application onto nodes of a distributed resource infrastructure, the determination
- 4 of the placement of the services onto the nodes comprising the steps of:
- 5 a. forming communication constraints between node pairs which ensure that
- 6 a sum of transport demands between a particular node pair does not exceed a
- 7 transport capacity between the particular node pair, each term of the sum
- 8 comprising a product of a first placement variable, a second placement
- 9 variable, and the transport demand between the services associated with the
- 10 first and second placement variables;
- 11 b. forming an objective; and
- 12 c. employing a local search solution to solve an integer program comprising
- 13 the communication constraints and the objective, which determines the
- 14 placement of the services onto the nodes.

1 22. A computer readable memory comprising computer code for directing a  
2 computer to make a determination of a placement of services of a distributed  
3 application onto nodes of a distributed resource infrastructure, the determination  
4 of the placement of the services onto the nodes comprising the steps of:  
5 a. establishing an application model of the services comprising transport  
6 demands between the services;  
7 b. establishing an infrastructure model of the nodes comprising transport  
8 capacities between the nodes;  
9 c. forming an integer program that comprises:  
10 i. a set of placement variables for a combination of the services and the  
11 nodes, each of the placement variables indicating whether a particular  
12 service is located on a particular node;  
13 ii. communication constraints between node pairs which ensure that a  
14 sum of the transport demands between a particular node pair does not  
15 exceed the transport capacity between the particular node pair, each term  
16 of the sum comprising a product of a first placement variable, a second  
17 placement variable, and the transport demand between the services  
18 associated with the first and second placement variables; and  
19 iii. an objective; and  
20 d. employing a local search solution to solve the integer program which  
21 determines the placement of the services onto the nodes.

1 23. The computer readable memory of claim 22 wherein the step of solving the  
2 integer program employs a local search solution.

1 24. The computer readable memory of claim 22 wherein the objective comprises  
2 minimizing communication traffic between the nodes.

1 25. The computer readable memory of claim 22 wherein the application model  
2 further comprises processing demands for the services.

1 26. The computer readable memory of claim 25 wherein the infrastructure model  
2 further comprises processing capacities for the nodes.

- 1 27. The computer readable memory of claim 26 wherein the integer program  
2 further comprises processing constraints ensure that a sum of the processing  
3 demands for each of the nodes does not exceed the processing capacity for the  
4 node.
- 1 28. The computer readable memory of claim 27 wherein the objective comprises  
2 balancing the processing demands on the nodes.
- 1 29. The computer readable memory of claim 26 wherein the processing demands  
2 and the processing capacities are normalized according to a processing criterion.
- 1 30. The computer readable memory of claim 29 wherein the processing criterion  
2 comprises an algorithm speed.
- 1 31. The computer readable memory of claim 9 wherein the processing criterion  
2 comprises a transaction speed.
- 1 32. The computer readable memory of claim 9 wherein the processing capacities  
2 of the nodes are found according to a look-up table in which different types of  
3 nodes have been normalized according to the processing criterion.
- 1 33. The computer readable memory of claim 22 wherein the application model  
2 further comprises storage demands for the services.
- 1 34. The computer readable memory of claim 33 wherein the infrastructure model  
2 further comprises storage capacities for the nodes.
- 1 35. The computer readable memory of claim 34 wherein the integer program  
2 further comprises storage constraints which ensure that a sum of the storage  
3 demands for each of the nodes does not exceed the storage capacity for the node.
- 1 36. The computer readable memory of claim 22 wherein the integer program  
2 further comprises placement constraints which ensure that each of the services is  
3 placed on one and only one of the nodes.

1 37. The computer readable memory of claim 22 wherein the services reside on the  
2 nodes according to a previous assignment.

1 38. The computer readable memory of claim 37 further comprising the step of  
2 assessing reassignment penalties for service placements that differs from the  
3 previous assignment.

1 39. The computer readable memory of claim 38 wherein the integer program  
2 further comprises a second objective that seeks to minimize the reassignment  
3 penalties.

1 40. A computer readable memory comprising computer code for directing a  
2 computer to make a determination of a placement of services of a distributed  
3 application onto nodes of a distributed resource infrastructure, the determination  
4 of the placement of the services onto the nodes comprising the steps of:  
5 a. establishing an application model of the services that comprises  
6 processing demands for the services, storage demands for the services, and  
7 transport demands between the services;  
8 b. establishing an infrastructure model of the nodes that comprises processing  
9 capacities for the nodes, storage capacities for the nodes, and transport  
10 capacities between the nodes;  
11 c. forming an integer program that comprises:  
12 i. a set of placement variables for a combination of the services and the  
13 nodes, each of the placement variables indicating whether a particular  
14 service is located on a particular node;  
15 ii. processing constraints which ensure that a sum of the processing  
16 demands for each of the nodes does not exceed the processing capacity for  
17 the node;  
18 iii. storage constraints which ensure that a sum of the storage demands for  
19 each of the nodes does not exceed the storage capacity for the node;  
20 iv. placement constraints which ensure that each of the services is placed  
21 on one and only one node;  
22 v. communication constraints between node pairs which ensure that a

23           sum of the transport demands between a particular node pair does not  
24           exceed the transport capacity between the particular node pair, each term  
25           of the sum comprising a product of a first placement variable, a second  
26           placement variable, and the transport demand between the services  
27           associated with the first and second placement variables; and  
28       vi.     an objective of minimizing communication traffic between the nodes  
29           and balancing processing loads on the nodes; and  
30       d.     employing a local search solution to solve the integer program which  
31           determines the placement of the services onto the nodes.